Bernoulli trials with variable probabilities - an observation by Feller

Peter Haggstrom
www.gotohaggstrom.com
mathsatbondibeach@gmail.com

September 26, 2023

1 Background

In his famous book, "An Introduction to Probability Theory and its Applications", Third Edition. Volume 1 [3] William Feller develops an interesting and somewhat counterintuitive result for the sum of n mutually independent random variables X_k such that each one assumes the values 1 and 0 with probabilities p_k and $q_k = 1 - p_k$ respectively (see [3] pages 230-231). He is interested in an interpretation of the variance of the sum $S_n = X_1 + X_2 + \cdots + X_n$ so he proceeds as follows.

For each k we have that the expectation of X_k, $E(X_k) = p_k$ and the variance:

$$\text{Var}(X_k) = E(X_k^2) - (E(X_k))^2 = p_k - p_k^2 = p_k q_k \tag{1}$$

If you have trouble seeing where these results come from, recall that since the Bernoulli variable X_k can take the value 1 with probability p_k and 0 with probability q_k, hence $E(X_k) = 1 \times p_k + 0 \times q_k = p_k$. The variance is defined as $\text{Var}(X_k) = E((X_k - \mu)^2) = E(X_k^2 - 2\mu X_k + \mu^2) = E(X_k^2) - 2\mu E(X_k) + E(\mu^2) = E(X_k^2) - 2\mu^2 + \mu^2 = E(X_k^2) - \mu^2$. Note that $E(X_k) = \mu$ and the linearity of the expectation operator has been used. Also the expectation of a constant, eg μ^2 is just the constant itself. Finally note that $E(X_k^2) = 1^2 \times p_k + 0^2 \times q_k = p_k$.

We know that the variance of the sum of n mutually independent random variables X_k, $S_n = \sum_{k=1}^{n} X_k$ is:

$$\text{Var}(S_n) = \sum_{k=1}^{n} \sigma_k^2 \quad \text{where: } \sigma_k = \sqrt{\text{Var}(X_k)} \tag{2}$$

A proof of (2) can be found in the Appendix.

Hence, using (1) we have:

$$\mathrm{Var}(S_n) = \sum_{k=1}^{n} \mathrm{Var}(X_k) = \sum_{k=1}^{n} p_k q_k \tag{3}$$

2 Feller's description of the issue

I will simply quote Feller at [3 page 231]:

"..the variable S_n may be interpreted as the total number of successes in n independent trials, each of which results in success or failure. Then $p = \frac{p_1+\cdots+p_n}{n}$ is the average probability of success, and it seems natural to compare the present situation to Bernoulli trials with the constant probability of success p. Such a comparison leads us to a striking result. We may rewrite (3) in the form:

$$\mathrm{Var}(S_n) = np - \sum_{k=1}^{n} p_k^2 \tag{4}$$

Next, it is easily seen (by elementary calculus or induction) that among all combinations p_k such that $\sum_{k=1}^{n} p_k = np$ the sum $\sum_{k=1}^{n} p_k^2$ assumes its minimum value when all the p_k are equal. It follows that, if the average probability of success p is kept constant, $\mathrm{Var}(S_n)$ *assumes its maximum value when* $p_1 = \cdots = p_n = p$. We have thus the surprising result that the *variability of* p_k*, or lack of uniformity, decreases the magnitude of chance fluctuations* as measured by the variance. For example, the number of fires in a community may be treated as a random variable; for a given average number, the variability is *maximal* if all households have the *same* probability of fire. Given a certain average quality p of n machines, the *output will be least uniform if all machines are equal.* (An application to modern education is obvious but hopeless.)"

3 Discussion

The final italicized comment by Feller above is indeed surprising even given the simplifying assumptions made. Notwithstanding Feller's stature as a probabilist (he was responsible for major analytic work on the Central Limit Theorem in the 1930s) let's nevertheless satisfy ourselves that he is right.

The way the issue is set up what we have is an optimisation problem:

Minimise $\sum_{k=1}^{n} p_k^2$ subject to the constraint: $p = \frac{p_1+\cdots+p_n}{n}$ is constant ie $\sum_{k=1}^{n} p_k = np$

In the language of Lagrange multipliers we thus have something of the form of minimising $F(p_1, p_2, \ldots, p_n)$ subject to $\phi(p_1, p_2, \ldots, p_n) = 0$. Here $F(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k^2$ and $\phi(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k - np = 0$. Thus we form the auxiliary function:

$$G(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda\phi(p_1, p_2, \ldots, p_n) \tag{5}$$

To find an extremum we need:

$$\frac{\partial G}{\partial p_k} = 0 \quad \forall k \tag{6}$$

Recall that this is a necessary condition so further investigation is needed to establish that we actually have a minimum.

Making the relevant substitutions in (5) we have:

$$G(p_1, p_2, \ldots, p_n) = \sum_{k=1}^{n} p_k^2 + \lambda\big(\sum_{k=1}^{n} p_k - np\big) \tag{7}$$

Differentiating we get:

$$\frac{\partial G}{\partial p_k} = 2p_k + \lambda = 0 \implies \lambda = -2p_k \quad \forall k \tag{8}$$

Now (8) only makes sense when the p_k are constant ie $p_k = p^*$ for all k. The constraint $\sum_{k=1}^{n} p_k - np = 0$ then becomes $\sum_{k=1}^{n} p^* - np = 0$. Therefore $np^* - np = 0$ and so $p^* = p = \frac{p_1 + \cdots + p_n}{n}$.

To see that $p = \frac{p_1 + \cdots + p_n}{n} = p_k \ \forall k$ actually minimises $\sum_{k=1}^{n} p_k^2$ perturb two of the p_k as follows. Without loss of generality we can relabel the p_k so that $p_1 \leq p_2 \leq \cdots \leq p_n$. Let:

$$p_1^* = p_1 - \epsilon, \quad p_2^* = p_2 + \epsilon \tag{9}$$

where $\epsilon > 0$. The other values of p_k remain the same ie $p_k^* = p_k$ for $k \neq 1, 2$. Hence $np = np^* = n\sum_{k=1}^{n} p_k^*$ ie it is constant. Thus:

$$\sum_{k=1}^{n} p_k^2 - \sum_{k=1}^{n} {p_k^*}^2 = \sum_{k=1}^{n} p_k^2 - \big((p_1 - \epsilon)^2 + (p_2 + \epsilon)^2\big) + p_3^2 + \cdots + p_n^2\big) = -2\epsilon^2 - 2\epsilon(p_2 - p_1) < 0 \tag{10}$$

Noting the assumption that $p_1 \leq p_2$. Thus $\sum_{k=1}^{n} p_k^2 < \sum_{k=1}^{n} {p_k^*}^2$ and so we do indeed have a minimum.

In two dimensions we can satisfy ourselves that $p = \frac{p_1+p_2}{2}$ minimises $\sum_{k=1}^{2} p_k^2$ and hence maximises the variance defined in (4) as $\mathrm{Var}(S_n) = np - \sum_{k=1}^{n} p_k^2$. Let:

$$f(p_1, p_2) = 2\big(\frac{p_1+p_2}{2}\big) - \sum_{k=1}^{2} p_k^2 = p_1 + p_2 - \sum_{k=1}^{2} p_k^2 \tag{11}$$

Recall that if $f : \mathbb{R}^2 \to \mathbb{R}$ is differentiable at $\vec{p}$ with continuous second partials at all points sufficienty close to $\vec{p}$ and $\vec{\nabla} f(\vec{p}) = \vec{0}$, then if the determinant of the Hessian at $\vec{p}$ is strictly positive then if $\frac{\partial^2 f}{\partial x^2} > 0$ at $\vec{p}$ implies a local minimum at $\vec{p}$.

The Hessian is defined as follows:

$$\boldsymbol{H(p_1, p_2)} = \begin{pmatrix} \frac{\partial^2 f}{\partial p_1^2} & \frac{\partial^2 f}{\partial p_2 \, \partial p_1} \\ \\ \frac{\partial^2 f}{\partial p_1 \partial p_2} & \frac{\partial^2 f}{\partial p_2^2} \end{pmatrix} \tag{12}$$

Of course the assumption of continuous second partials means that $\frac{\partial^2 f}{\partial p_2 \, \partial p_1} = \frac{\partial^2 f}{\partial p_1 \partial p_2}$ but (12) gives the structure in the general case.

Thus:

$$\det \boldsymbol{H(p_1, p_2)} = \begin{vmatrix} -2 & 0 \\ 0 & -2 \end{vmatrix} = 4 > 0 \tag{13}$$

Note that the extremum is given by:

$$\vec{\nabla} f(p_1, p_2) = \begin{pmatrix} \frac{\partial f}{\partial p_1} \\ \\ \frac{\partial f}{\partial p_2} \end{pmatrix} = \begin{pmatrix} 1 - 2p_1 \\ \\ 1 - 2p_2 \end{pmatrix} = \begin{pmatrix} 0 \\ \\ 0 \end{pmatrix} \implies p_1 = p_2 = \frac{1}{2} \tag{14}$$

Thus we have a local minimum at $p = \frac{p_1+p_2}{2} = \frac{1}{2}$.

In n dimensions the local minimum will exist where the Hessian is non-zero and the eigenvalues of the Hessian at the extremum are all positive. See [2 pages 260-262] for a discussion.

An inductive proof (alluded to by Feller) is perhaps more "satisfying" in showing that $p = \frac{p_1+\cdots+p_n}{n}$ minimises $\sum_{k=1}^{n} p_k^2$. The base case of $n = 1$ is trivial (and bereft of useful insight) so we consider $n = 2$. Thus:

$$\sum_{k=1}^{2} p_k^2 - \sum_{k=1}^{2} p^2 = p_1^2 + p_2^2 - 2\Big(\frac{p_1+p_2}{2}\Big)^2 = \frac{p_1^2 - 2p_1p_2 + p_2^2}{2} = \frac{(p_1-p_2)^2}{2} \geq 0 \tag{15}$$

Thus $\sum_{k=1}^{2} p^2$ is minimal. Now for $n+1$ we have, where $p^* = \frac{p_1+\cdots+p_n+p_{n+1}}{n+1}$:

$$\begin{aligned}
\sum_{k=1}^{n+1} p_k^2 - \sum_{k=1}^{n+1} p^{*\,2} &= \sum_{k=1}^{n+1} p_k^2 - (n+1)\frac{(p_1+\cdots+p_n+p_{n+1})^2}{(n+1)^2} \\
&= \sum_{k=1}^{n} p_k^2 + p_{n+1}^2 - \frac{(np+p_{n+1})^2}{n+1} \\
&\geq np^2 + p_{n+1}^2 - \frac{(n^2p^2 + 2n\,p\,p_{n+1} + p_{n+1}^2)}{n+1} \\
&\quad \text{using the induction hypothesis that } \sum_{k=1}^{n} p_k^2 \geq \sum_{k=1}^{n} p^2 \text{ and } np = \sum_{k=1}^{n} p_k \text{ is constant} \\
&= \frac{n(n+1)p^2 + (n+1)p_{n+1}^2 - n^2p^2 - 2n\,p\,p_{n+1} - p_{n+1}^2}{n+1} \\
&= \frac{n^2p^2 + np^2 + np_{n+1}^2 + p_{n+1}^2 - n^2p^2 - 2n\,p\,p_{n+1} - p_{n+1}^2}{n+1} \\
&= \frac{np^2 - 2n\,p\,p_{n+1} + np_{n+1}^2}{n+1} \\
&= \frac{n(p-p_{n+1})^2}{n+1} \geq 0
\end{aligned} \tag{16}$$

Thus $\sum_{k=1}^{n+1} p_k^2 \geq \sum_{k=1}^{n+1} p^{*\,2}$ and so the proposition is established by induction.

There is an analogy between this problem and that of finding the discrete probability distribution on the points $\{p_1, p_2, \ldots, p_n\}$ with maximal information entropy. Thus we need to maximise the Shannon entropy defined by:

$$f(p_1, \ldots, p_n) = -\sum_{k=1}^{n} p_k \log_2 p_k \tag{17}$$

For there to be a legitimate probability distribution we need the following constraint:

$$g(p_1, \ldots, p_n) = \sum_{k=1}^{n} p_k = 1 \tag{18}$$

Using Lagrange multipliers we proceed by forming the auxiliary equation:

$$f(p_1, \ldots, p_n) + \lambda\left(g(p_1, \ldots, p_n) - 1\right) \tag{19}$$

and so we require:

$$\frac{\partial}{\partial p_k}\Big\{ -\sum_{k=1}^{n} p_k \log_2 p_k + \lambda\Big(\sum_{k=1}^{n} p_k - 1\Big)\Big\} = 0 \quad \text{evalulated at } \vec{p} = \vec{p^*} \tag{20}$$

There are n separate equations in (20) and when we carry put the dfferentiation for $k = 1, 2, \ldots, n$ we get:

$$\begin{aligned} 0 &= -\log_2 p_k^* - p_k \frac{\partial(\log_2 p_k^*)}{\partial p_k} + \lambda \\ &= -\log_2 p_k^* - p_k \frac{\partial(\frac{\ln p_k^*}{\ln 2})}{\partial p_k} + \lambda \\ &= -(\log_2 p_k^* + \frac{1}{\ln 2}) + \lambda \\ \therefore \lambda &= \log_2 p_k^* + \frac{1}{\ln 2} \end{aligned} \tag{21}$$

The last line of (21) implies that all the p_k^* are equal because they depend solely on λ. Thus $p_k^* = p$ for all k but the constraint $\sum_{k=1}^{n} p_k^* = 1$ means that $np = 1$. Thus the required distribution is uniform with probability $p_k^* = \frac{1}{n}$.

More detail on optimisation theory applying to entropy maximisation can be found in [4, pages 222-228]

Feller also notes at [3, page 282] that S_n has a Poisson distribution in the limit ie where the p_k depend on n is such as way that the largest p_k tends to zero, but the sum $p_1 + p_2 + \cdots + p_n = \lambda$ remains constant. This result is derived using probability generating function techniques.

4 Application to investment and MOOCs

In the investment world each asset manager invests in a variety of assets eg domestic shares, international shares, bonds etc and their individual performance for each asset class can be measured. However, in so doing one has to be aware of the games that the asset managers play. For instance, they can pick benchmarks that are relatively easy to exceed. They can play with how they disclose performance eg pre-tax and fees versus after tax and fees. Frequently performance is related to the performance of the median manager for that asset class ie 50% of the managers will have performance higher than that manager. Feller's comment that "given a certain average quality p of n machines, the *output will be least uniform if all machines are equal* clearly has fundamental relevance to the performance of asset managers. If we found that the average probability of a universe of asset managers exceeding some asset class benchmark was 70%, say, the

performance of this group of managers would be least uniform if all the managers beat the benchmark with this probability. Given the propensity for herd like behaviour among asset managers in some contexts this theoretical proposition could assume more practical significance. I am not aware of an empirical work on this specific issue. This principle also has important implications for massive on-line open courses (MOOCS).

5 Appendix

To prove equation (2) we suppose that $X_1, \ldots, X_n$ are n mutually independent random variables with finite variances $\sigma_1^2, \ldots, \sigma_n^2$ and $S_n = X_1 + \cdots + X_n$.

We let $\mu_k = E[X_k]$ and $m_n = \mu_1 + \cdots + \mu_n = E[S_n]$. Then $S_n - m_n = \sum_{k=1}^{n}(X_k - \mu_k)$

Then:

$$\begin{aligned}(S_n - m_n)^2 &= \Big(\sum_{j=1}^{n}(X_j - \mu_j)\Big)\Big(\sum_{k=1}^{n}(X_k - \mu_k)\Big) \\ &= \sum_{k=1}^{n}(X_k - \mu_k)^2 + 2\sum_{j<k}^{n}(X_j - \mu_j)(X_k - \mu_k)\end{aligned} \tag{22}$$

Taking expectations of both sides and using the linear properties of the expectation operator we get:

$$\begin{aligned}&\mathrm{Var}(S_n) = E\Big[(S_n - m_n)^2\Big] = E\Big[\sum_{k=1}^{n}(X_k - \mu_k)^2 + 2\sum_{j<k}^{n}(X_j - \mu_j)(X_k - \mu_k)\Big] \\ &= E\Big[\sum_{k=1}^{n}(X_k - \mu_k)^2\Big] + 2E\Big[\sum_{j<k}^{n}(X_j - \mu_j)(X_k - \mu_k)\Big] \\ &= \sum_{k=1}^{n}E\Big[(X_k - \mu_k)^2\Big] + 2\sum_{j<k}^{n}\mathrm{Cov}(X_j, X_k) \\ &= \sum_{k=1}^{n}E\Big[(X_k - \mu_k)^2\Big] \\ &= \sum_{k=1}^{n}\sigma_k^2\end{aligned} \tag{23}$$

In (23) the term $2\sum_{j<k}^{n}\mathrm{Cov}(X_j, X_k) = 0$ because the X_j, X_k are mutually indepedent. This fundamental fact about the covariance is established as follows:

$$
\begin{aligned}
&\mathrm{Cov}(X_j, X_k) = E\Big[(X_j - \mu_j)(X_k - \mu_k)\Big] \\
&= E\Big[X_j X_k - X_j\,\mu_k - X_k\,\mu_j + \mu_j\,\mu_k\Big] \\
&= E\Big[X_j X_k\Big] - E\Big[X_j\,\mu_k\Big] - E\Big[X_k\,\mu_j\Big] + E\Big[\mu_j\,\mu_k\Big] \\
&= E\Big[X_j\Big]\,E\Big[X_k\Big] - \mu_k E\Big[X_j\Big] - \mu_j E\Big[X_k\Big] + \mu_j\,\mu_k \quad \text{using independence } E[XY] = E[X]\,E[Y] \\
&= \mu_j\mu_k - \mu_k\,\mu_j - \mu_j\,\mu_k + \mu_j\,\mu_k \\
&= 0
\end{aligned}
\tag{24}
$$

6 References

1. Stephen Boyd and Lieven Vandenberghe, Convex Optimization,Cambridge University Press, 2004

2. David M Bressoud, Second Year Calculus: From Celestial Mechanics to Special Relativity , Springer, 1991

3. William Feller, "An Introduction to Probability Theory and its Applications", Third Edition. Volume 1, Wiley.

4. Claude E. Shannon, (July–October 1948). "A Mathematical Theory of Communication" Bell System Technical Journal 27 (3): 379–423 `http://www3.alcatel-lucent.com/bstj/vol27-1948/articles/bstj27-3-379.pdf`

7 History

Created: 15 October 2014

26 September 2023: corrected typo in (8) and corrected reference to 2 dimensions in the discussion immediately before (11)